食源性疾病

基于知识图谱技术的食源性疾病风险防范技术研究及应用

马大燕，潘登，张朝正，吴永宁
（国家食品安全风险评估中心，国家卫生健康委员会食品安全风险评估重点实验室，北京 100021）

摘 要：目的 研究食源性疾病风险防范技术研究及应用。**方法** 以知识图谱为载体，通过对食品安全事故流行病学调查技术指南及互联网数据进行知识抽取与关系挖掘，获取常见食源性疾病致病因子与临床表现、潜伏期、易感人群、食品来源之间关联关系。**结果** 构建了食源性疾病知识图谱，包括390个节点和1 375条边，并在此基础上搭建了一个知识问答系统，实现指定食源性疾病致病因子的临床表现、易感人群、可能食物来源、生物标本采集要求等答案的自动获取。**结论** 本研究创新提出基于知识图谱技术的食源性疾病风险防范体系设计，构建的食源性疾病知识图谱有效解决致病因子相关检测关键数据字段定义不清，数据孤岛严重等问题。本文搭建的食源性疾病知识问答系统，对提升公众的食源性疾病风险认知，并纠正医护人员对食源性疾病相关的食品安全知识和操作行为具有重要的现实意义。

关键词：食源性疾病；致病因子；风险防范；知识图谱；知识问答系统
中图分类号：R155 **文献标识码：**A **文章编号：**1004-8456(2022)05-1035-06
DOI:10.13590/j.cjfh.2022.05.027

Research and application of foodborne disease risk prevention based on knowledge graph technology

MA Dayan, PAN Deng, ZHANG Chaozheng, WU Yongning
(National Center for Food Safety Risk Assessment, Key Laboratory of Food Safety Risk Assessment, Ministry of Health, Beijing 100021, China)

Abstract: Objective To study the research and application of foodborne disease risk prevention. **Methods** Taking the knowledge graph as the carrier, through knowledge extraction and relationship mining on the technical guide for epidemiological investigation of food safety accidents and internet data, the association between the pathogenic factors of common foodborne diseases and clinical manifestations, incubation period, susceptible population and food sources was deeply explored and obtained. **Results** The knowledge graph of foodborne diseases was constructed, including 390 nodes and 1 375 edges. On this basis, a knowledge question answering system was constructed to realize the automatic acquisition of answers to the clinical manifestations, susceptible populations, possible food sources and biological sample collection requirements of designated foodborne disease pathogenic factors. **Conclusion** This research innovatively proposed the design of foodborne disease risk prevention system based on knowledge graph technology. Meanwhile, the constructed foodborne disease knowledge graph can effectively solve unclear definition of key information and serious data islands in the related detection of pathogenic factors. The question answering system built on this basis was of great practical importance to improve the public's knowledge of foodborne disease risks and to correct the food safety knowledge and operational behaviors of health care professionals related to foodborne diseases.

Key words: Foodborne disease; pathogenic factors; risk prevention; knowledge graph; knowledge question answering system

食源性疾病是指人类由于摄入含病原物质的食品而引起的中毒或感染性疾病[1]，世界卫生组织（World Health Organization，WHO）报告显示，全球每年仅食源性或水源性腹泻导致的死亡人数可达220万[2]，食源性疾病已成为世界最突出的公共卫生问题之一[3]。据统计，我国每年食源性疾病的发病人数约15 984.1万人次[4]，可见食源性疾病严重威胁到人类的生命健康。据统计，大多数食源性疾病由个体的不健康饮食行为直接造成[5]，且常常受到公众食源性疾病知识和态度的影响。

目前，我国已建立了健全完善的食源性疾病监

收稿日期：2022-08-30
基金项目：国家重点研发计划(2020YFF0305005)
作者简介：马大燕 女 博士后 研究方向为人工智能技术在食品安全领域的应用 E-mail：madayan@cfsa.net.cn
通信作者：吴永宁 男 研究员 研究方向为化学污染监控技术、食品污染与人体健康关系的风险评估研究 E-mail：wuyongning@cfsa.net.cn

测系统，包括8 500余哨点医院网，疾病预防控制部门对发病人数在2人及以上或死亡1人及以上的食源性疾病进行监测并上报[6-7]。食源性疾病监测体系涉及患者就诊、样本采集、检验、事件上报等多个环节及部门，任一环节出现问题都会造成错报、漏报的情况。有数据显示，各国（地区）食源性疾病的监测数据中不明原因的疾病占比很高，达到10%以上。研究表明，提高医务人员食源性疾病知识知晓率是提升食源性疾病监测水平的有效途径之一[8]。

近年来，知识图谱作为一种高效、智能的知识组织手段，被广泛应用在智能语义搜索、问答系统以及公安、医疗、军事等行业。本研究选择以知识图谱作为知识载体，从提高食源性疾病的监测预警能力实际需求出发，深度挖掘常见致病因子的临床表现、潜伏期、易感人群、食品来源等关联关系，实现食源性疾病的图谱化，并在此基础上构建一个知识问答系统，完成基于自然语言的食源性疾病相关知识检索与推理。

1 材料与方法

知识图谱本质上是一种称为语义网络的知识库，旨在描述客观世界的概念、实体、事件及其间的关系。根据覆盖范围的不同，知识图谱可以区分为应用相对广泛的通用知识图谱和与属于某个特定领域的行业知识图谱。通用知识图谱覆盖范围广，注重横向广度，强调融合更多的实体，通常采用自底向上的构建方式，从开放链接数据（“信息”）中抽取出置信度高的实体，再逐层构建实体与实体之间的联系；行业知识图谱指向一个特定的垂直行业，注重纵向深度，具有丰富的实体属性和数据模式，通常采用自顶向下的构建方式，先定义好本体与数据模式，在抽取实体加入到知识库。知识图谱的构建遵循知识抽取、知识融合、知识加工、知识应用的基本流程。从海量结构化和非结构化数据中进行实体、关系、属性和事件的信息提取，通过本体和实体对齐、指代消解解决多种类型的数据冲突问题，完成知识融合。将知识存储到知识库中，最后进行进一步的知识推理和图谱应用，知识图谱的基本构建流程如图1所示。

知识图谱为人工智能任务提供算法支撑的典型应用主要包括智能问答、智能搜索和智能推荐、决策分析系统等，并在公安、医疗、军事、金融等多个行业落地应用。知识图谱的应用则相当广泛，但是几乎没有为食源性疾病风险感知、防范构建知识图谱的研究。主要原因是常见食源性疾病致病因子相关检测的关键数据字段定义不清，数据编码不统一，数据的数量和质量问题较多，数据孤岛现象严重，这给需要大量训练文本的自动化抽取技术带来很大的挑战，因此目前的研究需要从一些专业文档入手，建立统一的食源性疾病分类方式或专业本体，从而达食源性疾病风险防范的目的。

2 结果

2.1 食源性疾病知识图谱构建

2.1.1 数据来源

依据食品安全事故流行病学调查技术指南（2012年版），分析食品安全事故常见致病因子的临床表现、潜伏期及生物标本采集要求一览表，经过整理、筛选，其中71个常见食源性疾病致病因子文本文件可作为实验初始数据集。以此为基础，从互联网上获取近5年内共1 268项食源性疾病暴发疾病数据样本，收集获得常见食源性疾病易感人群、食品来源等数据信息。

2.1.2 本体层建模

本体是对概念进行建模的规范，是描述客观世界的抽象模型，以形式化方式对概念及其之间的联系给出明确定义。本体可通过人工编辑方式手动构建，也可通过数据驱动自动构建。本研究采用自顶

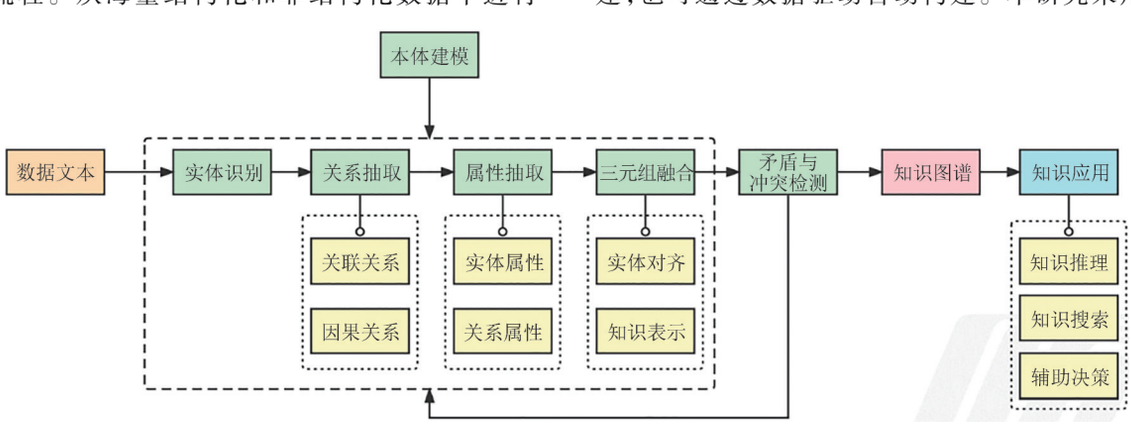


图1 知识图谱的基本构建流程

Figure 1 The built processing of knowledge graph

而下方式构建知识图谱，本体层则采用人工方式构建，后续则基于本体框架进行知识抽取；领域知识图谱通过抽象后得到标准的实体类型、关系和属性，即本体层模型，如图 2 所示。该本体层包括致病因子（含层级）、临床表现、易感人群、食品来源（含层级）、生物标本 5 种实体类型关系，实体类型和实体类型关系又包括 id、name、extn、feature、frequency、level 等共计 12 种属性。以食源性疾病知识图谱为例，致病因子（CausalFactor 实体类型）—表现（has_symptom 关系）—症状（Symptom 实体类型）、致病因子（CausalFactor 实体类型）—名称（name 属性）—诺如病毒（实体），实体类型有属性，关系也可以有属性，如 has_symptom 关系属性包括 dis_extn、extn、feature、weight_level 等。实体类型关系和属性的释义见表 1。

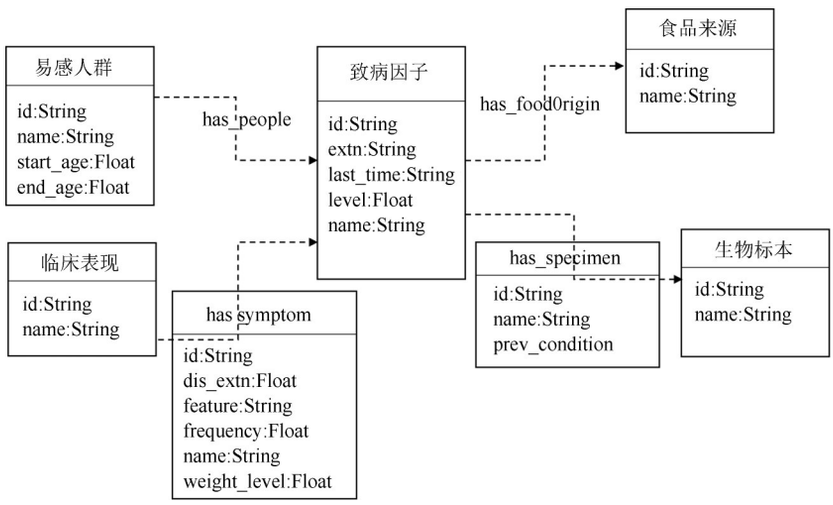


图 2　食源性疾病图谱的本体层

Figure 2　The ontology layer of foodborne disease knowledge graph

表 1　部分食源性疾病实体类型、关系及属性释义

Table 1　Definition of some entity type, relationship and attribute of foodborne disease knowledge graph

属性/关系编码		释义
id	属性&关系	知识元的唯一标识
name	属性&关系	知识元的名称
start_age	属性	易感人群—归属于—起始年龄
end_age	属性	易感人群—归属于—终止年龄
last_time	属性	致病因子—存在—潜伏期
level	属性	1 和 2，致病因子—归属于—1 级或致病因子—归属于—2 级
has_symptom	关系	致病因子—表现为—临床表现
has_people	关系	致病因子—易感—人群
has_foodOrigin	关系	致病因子—来源—食品
has_specimen	关系	致病因子—采集—生物标本
causal_Factor_hierachy	关系	致病因子层级—归属于——致病因子层级
food_Origin_hierachy	关系	食品层级—归属于——食品层级

2.1.3　实例层建模

食源性疾病知识图谱采用自顶向下的构建方法，本体层（模式层）构建完成后，接下来进行实例层填充。基于结构化、半结构化和非结构化的信息资源，通过知识抽取技术，从这些不同结构和类型的数据中提取出计算机可理解和计算的结构化数据；综合考虑训练资源缺乏、食源性疾病电子病例、指南等数量繁多，且标准间的形式与定义规则皆不统一等多方面因素，采用规则和监督学习相结合的实体抽取和关系抽取方法完成知识图谱知识获取工作。

2.1.3.1　实体类型关系/属性三元组抽取

食品安全事故流行病学调查技术指南（2012 年版）中常见致病因子与临床表现、潜伏期及生物标本采集要求等关系较为常见，主要进行这些实体类型对应三元组的提取，提取方式是基于规则与正则表达式的匹配。实体类型关系/属性三元组（h，r，t）中头实体（h）均对应食源性疾病致病因子，三元组提取的关键是尾实体（t）的提取，潜伏期是由数字+时量词（年、月、天、时、分钟、秒等）组成，可以采用正则表达式准确提取到致病因子对应的潜伏期属性。临床表现实体类型则基于临床医学标准术语词表，采用早期（症状）/晚期（症状）+临床表现术语的方式进行临床表现尾实体的提取。需要注意的是，部分症状除名称属性外，还包括性质、程度、发生条件等属性。以发生条件属性为例，常见的有“晨起加重”“行走时加重”等，这些词汇在医疗系统属于通用词汇，但缺乏标准化、统一的标准术语表，需要借助行业专家进行数据校对，得到完备的实体类型属性及关系属性三元组。

2.1.3.2　知识融合

由于知识来源广泛，存在知识质量良莠不齐、不同数据源的知识重复、层次结构缺失的问题，所以有必要基于合理的知识表示方法进行知识的融合。知识融合使来自不同知识源的知识在同一框

架规范下进行异构数据整合、加工、知识更新与验证等步骤，达到知识图谱在模式层和数据层的统一和规范。本研究涉及到的关键技术是实体对齐，解决来自不同文件的实体表达形式不一致的问题，如发芽马铃薯和龙葵素意思相近，但表达形式不一致，这时需要用到实体对齐。目前实体对齐最常用的方法是embedding[9]方法，由于食源性疾病相关文本训练数据集不足，且涉及到的致病因子潜伏期、分层等属性单一，无法基于embedding方法进行训练。本文通过构建同义词库+MinHashLSH文本相似度计算来解决这个问题，对基于同义词库未完成匹配的实体，通过MinHashLSH算法（设threshold=0.4）进行文本相似度计算，threshold阈值可根据图谱知识质量进行调整，最后借助行业专家完成实体的对齐及校验。

2.1.3.3 Neo4j图数据库扩充

从互联网上获取近5年内食源性疾病暴发事件样本数据，针对网页半结构化的数据形式，通过网络爬虫+模板对网络数据进行预处理，收集获得常见食源性疾病对应易感人群，食品来源等数据信息，将处理结果同样经过2.1.3.1和2.1.3.2，实现满足知识质量前提下图谱的知识扩充。

Neo4j图数据库具有实时数据分析、轻松检索、高可用性等优点，并且可以可视化显示节点间关联关系，数据库查询速度快，代码量少。因此，通过行业专家校验的知识三元组采用Neo4j作为存储数据库。基于Neo4j图数据库构建的食源性疾病知识图谱可视化界面如图3所示。

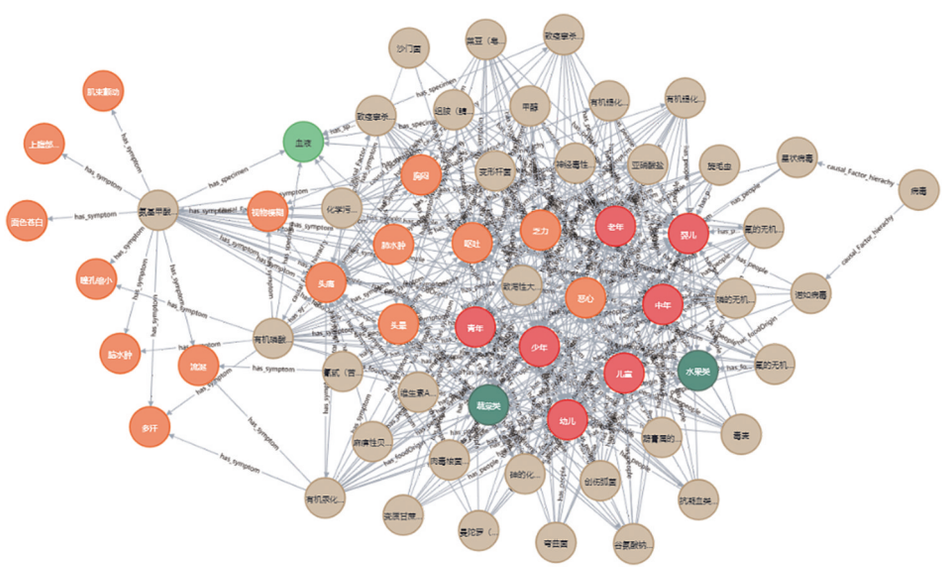


图3 食源性疾病知识图谱部分知识三元组可视化界面

Figure 3 The visual interface of partial knowledge triad of foodborne disease knowledge graph

2.2 食源性疾病知识问答系统搭建

本研究通过使用Rasa的机器对话框架和Neo4j图数据库，构造了一个食源性疾病知识问答系统，以便通过聊天的方式精准地获取到指定食源性疾病致病因子的临床表现、易感人群、可能食物来源、生物标本采集要求等信息。Rasa支持Spacy、MITIE、Jieba等多种开源NLP工具，构造定制化的NLP Pipeline达到对输入文本的自然语言语义理解及对话管理，通过与食源性疾病知识图谱建立联系，实现根据构造问题动态获取数据、构造答案、返回答案的目的。NLP Pipeline工作流程如图4所示。

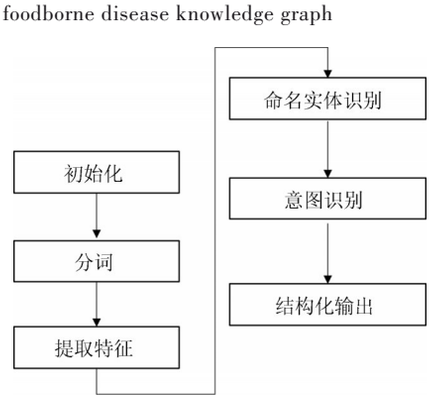


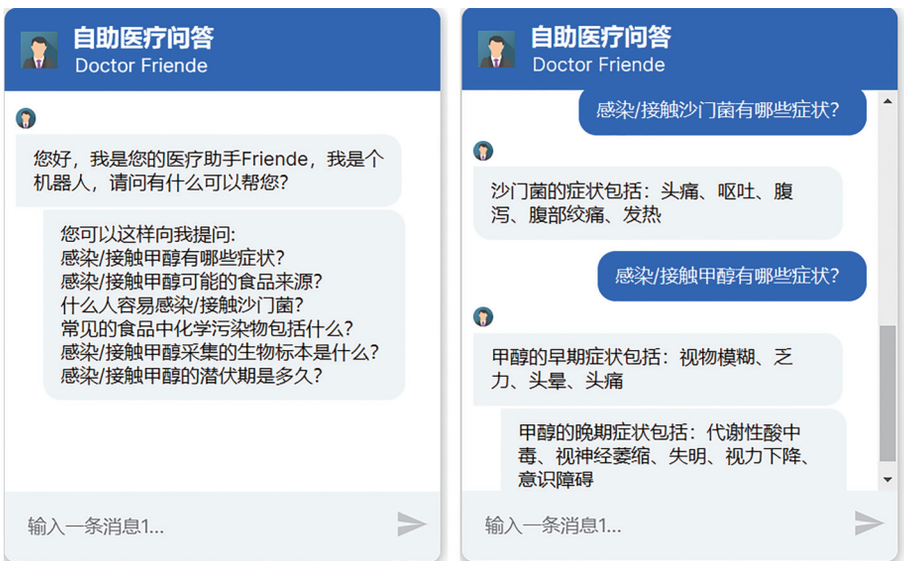
图4 Rasa NLP Pipeline工作流程

Figure 4 The workflow of Rasa NLP pipeline

本研究知识问答系统共包括六大功能模块，分别是“查询感染/接触致病因子有什么临床表现？”“查询感染/接触致病因子可能吃了什么？”“查询什么人容易感染/接触致病因子？”“查询常见的食品中致病因子包括哪些？”“查询常见致病因子的生物标本采集要求？”“询感染/接触常见的致病因子的潜伏期是多久？”。

以查询感染/接触致病因子有什么临床表现功能模块时，首先用户向系统提出问题，系统对原始

问题进行分词和关键词提取，然后对抽象化后的句子进行意图识别，成功匹配到“action_search_symptom”这一意图，通过访问 Neo4j 图数据库返回用户需要的答案。系统前端展示如图 5 所示。



(a) 初始化系统界面　　(b) 查询致病因子临床表现界面

图 5　知识问答系统前端效果图

Figure 5　The front end interface of knowledge question answering system

对超出知识问答系统数据库知识范围的问题，如食源性疾病知识库中并没有大肠菌群这条数据，当用户提问与该条数据相关的问题时，系统则在正确识别提问意图后，返回“知识库中暂无与大肠菌群相关的记录”，并将对话流程转换到下一轮问答中，系统前端展示界面如图 6 所示。

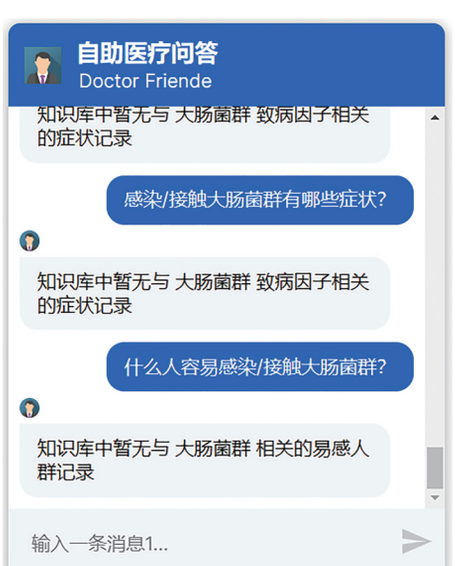


图 6　查询超出数据库范围前端效果图

Figure 6　Front end interface of query beyond database range

3　结论

本文通过对现有搜集到的食源性疾病知识进行实体与关系抽取，构建了食源性疾病知识图谱，整个图谱包括 390 个节点和 1 375 条边。Neo4j 图数据库作为知识三元组的存储方式，具备灵活多样、轻松检索、拓展性强等特性，通过输入简单的 Cypher 查询语言可实现 Neo4j 图数据库节点/关系的查询及可视化展示。图谱构建过程中完成了各实体类型的实体对齐，这对未来食源性疾病图谱构建采用更自动化的知识抽取提供了基础与依据。

智能问答作为知识图谱的应用之一被广泛应用于远程医疗问诊、智能客服、手机智能助手等应用场景中。本文在充分分析食源性疾病知识结构和风险防范关键基础上，设计了支持风险感知和防范的食源性疾病知识图谱工作流程。采用 Rasa 的机器对话框架，连接 Neo4j 图数据库，构造了一个食源性疾病知识问答系统，方便用户以聊天的方式精准获取常见食源性疾病致病因子的临床表现、易感人群、可能食物来源、生物标本采集要求等信息。本研究对有效干预和引导公众的食源性疾病风险感知偏差，并纠正医护人员对食源性疾病相关的食品安全知识和操作行为具有重要的现实意义。

参考文献

[1] 赵彤. 食源性致病菌检测现状与食品微生物危险性评估[J]. 中国卫生标准管理，2019，10(4)：7-9.
ZHAO T. Detection status of foodborne pathogenic bacteria and food microbial risk assessment [J]. China Health Standard Management，2019，10(4)：7-9.

[2] FOOD STANDARDS AGENCY. Foodborne disease strategy

2010-15: Objectives, vision and approach[S]. London: Food Standards Agency, 2011.

[3] 陆姣, 吴林海. 中国食源性疾病的风险特征研究[M]. 北京: 社会科学文献出版社·经济与管理分社, 2018: 7.
LU J, WU L H. A research on risk characteristic of foodborne disease in China [M]. Beijing: Social Sciences Literature Publishing House, 2018: 7.

[4] 毛雪丹, 胡俊峰, 刘秀梅. 我国细菌性食源性疾病疾病负担的初步研究[J]. 中国食品卫生杂志, 2011, 23(2): 132-136.
MAO X D, HU J F, LIU X M. Epidemiological burden of bacterial foodborne diseases in China—Preliminary study [J]. Chinese Journal of Food Hygiene, 2011, 23(2): 132-136.

[5] CHASSY B M. Food safety risks and consumer health[J]. New Biotechnology, 2010, 27(5): 534-544.

[6] 梁红云. 基层疾控机构食源性疾病暴发事件处置策略的探讨[J]. 中国药物与临床, 2018, 18(8): 1404-1405.
LIANG H Y. Discussion on handling strategies of foodborne disease outbreaks in grass-roots disease control institution[J]. Chinese Remedies & Clinics, 2018, 18(8): 1404-1405.

[7] 白莉, 刘继开, 李薇薇, 等. 中美食源性疾病监测体系比较研究[J]. 首都公共卫生, 2018, 12(2): 62-67.
BAI L, LIU J K, LI W W, et al. Comparison of foodborne disease surveillance systems in China and the US[J]. Capital Journal of Public Health, 2018, 12(2): 62-67.

[8] 苏涛, 毛永杨, 李智高, 等. 国内外食源性疾病监测与负担估计的研究进展[J]. 食品安全质量检测学报, 2019, 10(17): 5940-5946.
SU T, MAO Y Y, LI Z G, et al. Research progress on foodborne disease surveillance and burden estimation at home and abroad [J]. Journal of Food Safety & Quality, 2019, 10(17): 5940-5946.

[9] GUAN S P, JIN X L, WANG Y Z, et al. Self-learning and embedding based entity alignment[J]. Knowledge and Information Systems, 2019, 59(2): 361-386.